IGBONER 2.0: EXPANDING NAMED ENTITY RECOGNITION DATASETS VIA PROJECTION

Chiamaka Chukwuneke[1,2], **Paul Rayson**[1], **Ignatius Ezeani**[1,2], **Mahmoud El-Haj**[1]
Doris Asogwa[2], **Chidimma Okpalla**[3], **and Chinedu Mbonu**[2]

[1]Lancaster University, United Kingdom, [2]Nnamdi Azikiwe University, Nigeria,
[3]Federal University of Technology Owerri, Nigeria

ABSTRACT

Since the inception of the state-of-the-art neural network models for natural language processing research, the major challenge faced by low-resource languages is the lack or insufficiency of annotated training data. The named entity recognition (NER) task is no exception. The need for an efficient data creation and annotation process, especially for low-resource languages cannot be over-emphasized. In this work, we leverage an existing NER tool for English in a cross-language projection method that automatically creates a mapping dictionary of entities in a source language and their translations in the target language using a parallel English-Igbo corpus. The resultant mapping dictionary, which was manually checked and corrected by human annotators, was used to automatically generate and format an NER training dataset from the Igbo monolingual corpus thereby saving a lot of annotation time for the Igbo NER task. The generated dataset was also included in the training process and our experiments show improved performance results from previous works.

1 INTRODUCTION

Research shows that pre-trained language models achieve improved results for many natural language processing (NLP) tasks like Machine Translation (Nekoto et al., 2020), Topic Classification (Hedderich et al., 2020), Part of Speech Tagging (Kann et al., 2020) and Named Entity Recognition (NER) (Adelani et al., 2021; Ogueji et al., 2021). However, these models are known to require a lot of labelled data to perform well (Xie et al., 2020). For NLP tasks, there is a crucial and constant need for annotated corpora, which is often a challenge for low-resource languages. The African language Igbo is low-resourced for NLP research in general, including for the NER task (Adelani et al., 2021). This is a limitation for the IgboNLP research and in the training and evaluation of some state-of-the-art (SOTA) models built in other languages. This is portrayed in the evaluation result of the study by Chukwuneke et al. (2022) using IgboNER[1] dataset. IgboNER was created from only 3,190 sentences obtained from BBC Igbo news[2]. This forms our motivation in creating further annotated datasets for the low-resourced Igbo language. The crucial question then becomes how can we efficiently create more annotated data for Igbo NLP tasks?

In this work, we adopt the cross-language projection method ((Li et al., 2021); (Ehrmann et al., 2011); (Xie et al., 2018); (Bari et al., 2020)) to automatically create more NER datasets as an alternative to the time-consuming and costly human-annotated method. The projection-based method generates a tagged corpus by projecting the tags from the source language to the target language. A parallel corpus is usually required for this method. This study aims at creating more annotated NER data for the Igbo language using the projection-based method. Figure 1 is an example of a projection-based task. Kulshreshtha et al. (2020) showed that using parallel corpora is better suited for aligning contextual embeddings as the context of words, especially the ambiguous words are maintained within the sentences. English-Igbo parallel corpora were used in this work. Firstly, we

[1] https://github.com/Chiamakac/IgboNER-Models/tree/main/Igbo_dataset
[2] https://www.bbc.com/igbo

chose the English Language because it is the lingua franca of Nigeria where Igbo is one of the official native languages. Secondly, English has been widely researched in NLP and this has produced a lot of NLP resources for English. We tagged the English sentences (source language) using an existing English annotation tool and then projected the tags onto the parallel Igbo sentences (target language). The Igbo language has historically faced a lot of disagreement with the adoption of an official written orthography (Oraka, 1983). This greatly affected the language and resulted in Igbo not having a long written tradition when compared to languages like Arabic, English and French (Agbo, 2013). This has resulted in the scarcity of Igbo corpora online, and the few Igbo texts written with mixed orthographies. Also, the time, labour and resource needed to manually annotate a large gold standard corpus is a severe constraint.
The contributions of this paper include the creation of an additional NER dataset for Igbo Language via annotation projection using parallel corpora to augment the already existing IgboNER datasets. We also built a mapping dictionary containing all the unique entities identified by spaCy[3] with their tags. This to the best of our knowledge is the first IgboNER mapping dictionary, which is an important contribution to Igbo NLP and African NLP at large. Crucially, the dictionary was used to handle the issue of multi-words and spelling variation during the tag projection. Our experiments showed that there is an increase in model performance with increasing data size. The rest of the paper is organised as follows: Section 2 discusses the related work; Section 3 describes the data collection and annotation; Section 4 is about building the mapping dictionary; Section 5 is conclusion and future work.

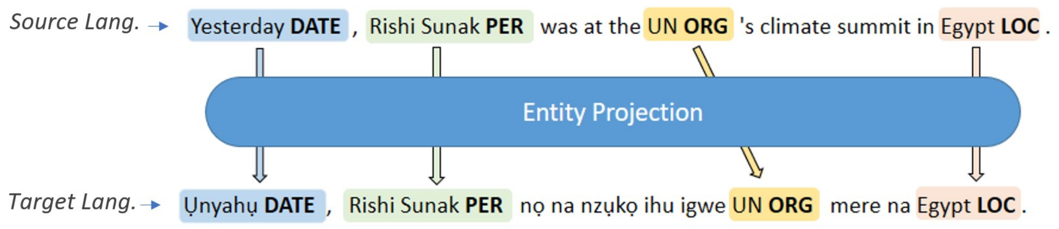
Source Lang.
Yesterday DATE , Rishi Sunak PER was at the UN ORG 's climate summit in Egypt LOC .
Entity Projection
Target Lang.
Ụnyahụ DATE , Rishi Sunak PER nọ na nzụkọ ihu igwe UN ORG mere na Egypt LOC .

Figure 1: Projecting the tags from the source language to the target language.

2 Related Work

Named Entity Recognition is a term defined as the task of identifying names of organisations, people, currency, time, percentage expression and geographic locations in text and was introduced at the Sixth Message Understanding Conference (Grishman & Sundheim, 1995). With the data-hungry deep neural networks recently used in training the SOTA models for performing NLP tasks as NER, the need for ever larger annotated datasets in various languages is inevitable. Over the past few years, attention has been drawn to weak or distant supervision[4] and a considerable number of studies (Mintz et al. (2009); Xiao et al. (2020); Liang et al. (2020); Hedderich et al. (2021); Rijhwani et al. (2020); Quirk & Poon (2017); Meng et al. (2021); Hogan (2022)) applied this method to generate a large amount of auto-labelled data without human effort. This method has helped address the issue of data scarcity and showed improvements in performance as seen in the above studies however, the availability of external resources like an existing knowledge base, gazetteers or dictionaries of named entities is a problem for some low-resource languages like Igbo.

The work by Li et al. Li et al. (2021) built an entity alignment model on top of an XLM-R (Conneau et al., 2020a) and projected the labelled entities on the source English language sentences of the parallel data to the target language sentences. They evaluated on four languages: Spanish, Chinese, German and Dutch. Their result showed Chinese to have the highest improvement in F1 score

[3] https://spacy.io/models/en_core_web_sm

[4] an automatic method of creating labelled data using an external source like an already existing knowledge base, gazetteers or dictionaries of named entities, etc

Type	Sentence pairs	Source
Igbo	5,503	BBC News
English	5,630	Local newspapers
Total	11,133	

Table 1: Data source. This describes the parallel data sources.

by 6.4%. Fei et al. (2020) implemented automatic corpus translation of the source gold-standard corpus to the target languages and then projected the labels on the source dataset to the translated target languages. Their evaluation result on the Universal Proposition Bank (UPB, v1.0)[5] dataset showed the translation-based method to be effective with an F1 score increase of 6.7%. Guo Guo & Roth (2021) translated high-resource labelled sentences to the target language word-by-word with a dictionary. Then, they constructed target-language text from the source-language named entities with a pre-trained language model. The study achieved state-of-the-art performances on LORELEI (low-resource) languages and perform comparatively on CoNLL (high-resource) languages. Garcia et al. García-Ferrero et al. (2022) automatically created data for the target language by translating the gold-labelled English data and then projected the gold labels from the source sentences to the translated sentences by leveraging automatic word alignments. Their result showed that data-based cross-lingual transfer approaches remain a useful option when high-capacity multilingual language models are not available. Enghoff et al. Enghoff et al. (2018) extensively studied annotation projection from multiple sources for low-resource NER. They showed that standalone multi-source annotation projection for NER works when the quality and availability of resources are high but suffers when parallel corpora are not available. THE PROJECTOR, an interactive graphical user interface tool that displays the sentence pair with all predicted and projected annotations was designed by Akbik & Vollgraf (2017). This is facilitate experiments and discussions in the research community. The study by Agerri et al. (2018) demonstrated the feasibility of automatically generating Named Entity Recognition (NER) taggers for a given language when no manual data is available. This they achieved with the use of parallel corpora, projecting the existing annotations from multiple source languages to the target language via strict match projections. Their evaluation showed that the automatically generated model outperforms the gold-standard trained model in an in-domain evaluation.

3 Data Collection and Annotation

3.1 Data collection

Our work used a parallel corpus (English-Igbo) that was originally created for a machine translation project (Ezeani et al., 2020). The corpus was created by collecting English and Igbo sentences from local newspapers in Nigeria (e.g. Punch) and from BBC Igbo News website[6] respectively. The 5,630 English sentences collected were translated into Igbo and the 5,503 collected Igbo sentences to English via human translation to produce English-Igbo sentence pairs used to build a standard machine translation benchmark dataset for Igbo. We adopted and used this specific corpus because it contains more contemporary texts in both languages which will guarantee a significant representation of known named entities than the multilingual bible text data[7] often used in the parallel corpus research. In addition, it was assumed that the quality of the data will be good enough due to the human-translation process that it has gone through. Table 1 presents a description of the data splits that make up the parallel corpus used in this work.

3.2 Data annotation

The key to building the mapping dictionary in the next section is the ability to identify all unique named entities in the source language text. Therefore, the initial data annotation (i.e. NER tagging) process involved running a NER tool over the source language (English in this case) to identify and

[5] https://github.com/System-T/UniversalPropositions

[6] https://www.bbc.com/igbo

[7] Mostly from https://www.jw.org and other sources

extract the named entities in the text and their tags. To achieve this, the English SpaCy[8] pipeline was used to extract the named entities in the English text. The pipeline we used is the spaCy model ‘en_core_web_sm version 3.4.0‘ [9] and the following tags or named entities were recognised by the model; PERSON(PER), NORP (Nationalities or religious or political groups), FAC (Faculty) ORG (Organisation), GPE (Geopolitical entity), LOC (Location), PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL. We passed our 11,133 English source sentences through SpaCy and it identified named entities in 7,894 sentences.This means that no entity was identified in 3,239 sentences. Table 2 gives a description of our annotated data.

Entity Types	**# Tagged entities**	**Cohen’s Kappa**	**Disagreement(%)**
PER	6,466	0.9971	0.3438
LOC	3,927	0.9968	0.2677
ORG	4,626	0.9821	2.4584
DATE	3,869	0.9966	3.2090

Table 2: Describes our dataset annotation including the entity types, number of tagged entities, Cohen’s Kappa inter-annotation scores and percentage disagreement.

4 Building the Mapping Dictionary

The mapping dictionary is a key contribution of this work and could be useful for other NLP tasks such as machine translation and other structured prediction systems. Its creation follows a semi-automatic process described in Figure 2. The key idea is based on the assumption that the majority (not all) of the named entities in the source language are going to either remain the same or translate to unique words or expressions in the target language. For example in Figure 1, `Rishi Sunak`, tagged PER, is always going to remain the same in both languages as it is a person’s name while `Yesterday`, tagged DATE, is most likely going to be translated as `Ụnyahụ`. For this work, the mapping dictionary is an actual python dictionary object which has a `key:value` structure. The entities from the English sentences are the *keys* while the corresponding Igbo entities and tags are the *values*. For example, looking at the sentences from Figure 1, the entry to the dictionary may be as shown below:

```
{
"Yesterday": {ig:"Ụnyahụ", tag:"DATE"}, "Rishi Sunak": {ig:"Rishi
Sunak", tag:"PER"},      "UN":  {ig:"UN", tag:"ORG"},      "Egypt"
{ig:"Egypt", tag:"LOC"},
}
```

As shown in Figure 2, the input to the process of building the mapping dictionary is the parallel corpus containing the sentence pairs in both languages as well as an initial mapping dictionary. The initial dictionary could be empty at the start of the process but may also have been initialised by a previous run. The English sentences are passed through the spaCy pipeline as described in section 3.2 to extract the NER-tagged entities. Each of the extracted entities is passed through the pipeline checking the following conditions:

- if the entity is already in the mapping dictionary, discard and get another entity.
- else if the entity is found in the target (Igbo) text, key = entity, value = entity, tag
- else if the entity’s translation is found in the target (Igbo) text, key = entity, value = translated, tag
- else manually annotate the target (Igbo) entity, key = entity, value = annotated, tag

Our experiment with the parallel text used shows that over 95% of the extracted entities from the source language (English) were found in the corresponding target (Igbo) sentence making human translation and annotation a lot easier.

[8] SpaCy is a free open-source library for Natural Language Processing in Python. SpaCy features NER tagging, part-of-speech tagging, dependency parsing, word vectors and more: https://spacy.io/

[9] Trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities

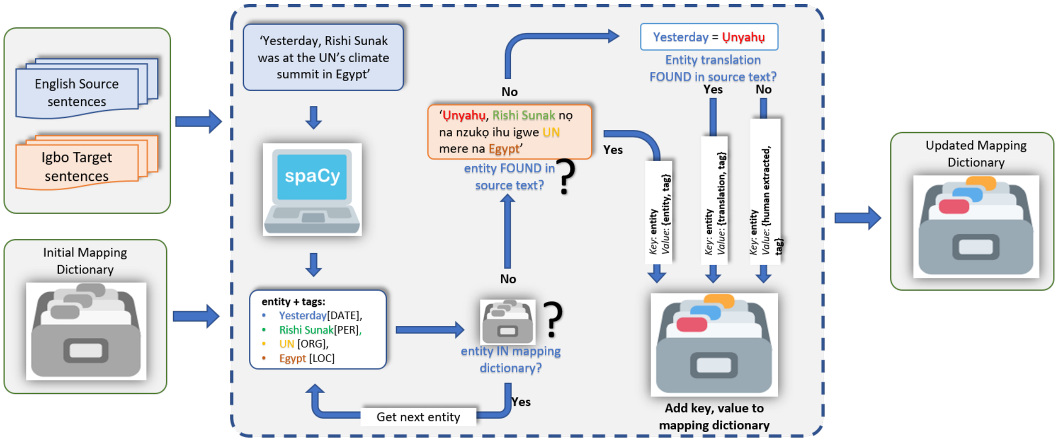
English Source sentences
Igbo Target sentences
Initial Mapping Dictionary
'Yesterday, Rishi Sunak was at the UN's climate summit in Egypt'
spaCy
entity + tags:
Yesterday[DATE],
Rishi Sunak[PER],
UN [ORG],
Egypt [LOC]
entity IN mapping dictionary?
Yes
Get next entity
No
'Ụnyahụ, Rishi Sunak nọ na nzukọ ihu igwe UN mere na Egypt'
entity FOUND in source text?
Yes
No
Yesterday = Ụnyahụ
Entity translation FOUND in source text?
Yes
No
Key: entity Value: {entity, tag}
Key: entity Value: {translation, tag}
Key: entity Value: {human extracted, tag}
Add key, value to mapping dictionary
Updated Mapping Dictionary

Figure 2: An illustration of the semi-automatic process used in creating the mapping dictionary.

4.1 Annotation Validation

Table 1 showed that our source data is from local news and newspapers which represents the daily communication of speakers of Igbo language. Because of the nature of our data (local news), spaCy's output was noisy. For example, an entity 'Sowore' was tagged as an ORG instead of a PERSON. The tagged spaCy output was therefore manually corrected by two people who are among the authors of this paper following the MUC-6 annotation guide[10] which we adopt in this work. Agreement was made by them where there were discrepancies in the tags. The inter-annotation agreement was calculated using Cohen's Kappa (Cohen, 1960) which uses one distribution for each rater and takes into account the possibility of the agreement occurring by chance. The inter-annotation percentage disagreement after the manual correction is in Table 1. Our dataset achieves an inter-annotator agreement of 0.985, i.e. excellent agreement. In section 3.2, we have the list of output tags by spaCy but for this work we are using only the PER, LOC, ORG, DATE tags for our dataset creation. We choose these four tags and replace all the 'GPE' tags with 'LOC', 'TIME' tag we also replace with 'DATE' following previously created IgboNER dataset (Chukwuneke et al., 2022). A python script which selects only the entity to replace was written and used for all the manual corrections. We use the annotated dataset as train split only, this is to increase the amount of training data for this work. As mentioned in section 3, our method requires a parallel corpus and we used the data described in Table 1.

4.2 Experimental Settings

In this section, we describe the IgboBERT 2.0 model we pre-trained for Igbo language and some state-of-the-art transformer models fine-tuned to the downstream Igbo NER task. We show the effect of dataset sizes on the performance by fine-tuning using various dataset sizes.

4.2.1 IgboBERT 2.0 Model

IgboBERT 2.0 is another version of IgboBERT (Chukwuneke et al., 2022) language model, which we pre-trained using the same corpus as IgboBERT plus the corpus described in Table 1 as an addition to increase the training sentences. A total number of 402,582 training sentences is used. We train IgboBERT 2.0 at a learning-rate of 1e-4 for 5 epoch and a batch size of 16. Only 5 epoch training was carried out because of limited compute resource such as GPU at the time of the experiment.

4.2.2 Fine-tuned Models

We fine-tune the IgboBERT 2.0 model and some of the state-of-the-art (SOTA) transformer models on the IgboNER downstream task using the Igbo dataset created by Chukwuneke et al. (2022),

[10] https://cs.nyu.edu/grishman/muc6.html

MasakhaNER 2.0 dataset (Adelani et al., 2022) and the dataset created in this work. The fine-tuned SOTA models are:

- Multilingual BERT-base cased (mBERT-base cased) by Devlin et al. (2019).
- XLM-RoBERTa-base (XLM-R) by Conneau et al. (2020b).
- DistilBERT by SANH et al.

To predict the entity types, we added a linear classification layer to the pre-trained transformer models. 30 epoch training with a batch size of 8 at a learning rate of 1e-4 was run.

4.3 Results

Table 3 shows the fine-tuned results of mBERT, XML-R, DistilBERT, and IgboBERT 2.0. We combine the ’train’ split of the Igbo dataset created by Chukwuneke et al. (2022), MasakhaNER 2.0 dataset (Adelani et al., 2022) and the dataset we created in this work. We divide the combination of the ’train’ split into 4 different data sizes: 25% (5726), 33% (7635), 50% (11,452) and 100% (22,904). We then fine-tune the above listed models with the 4 different data sizes. We validate and test the models with the same validation and test set. The comparitive evaluation shows:

- increase in performance with increase in the data size.
- XML-R outperformed others with F1-score of 88.83.
- IgboBERT 2.0 with F1 of 79.25 with datasize 22,904 is relatively comparable to the SOTA models in terms of the amount of data used in training these SOTA models.

		Percentage Data Sizes			
Models	**Scores**	**(25%)**	**(33%)**	**(50%)**	**(100%)**
IgBERT 2.0	F1	72.33	76.00	77.10	79.91
	Precision	71.49	74.29	75.93	79.25
	Recall	73.20	77.79	78.31	80.58
mBERT	F1	80.93	82.64	83.58	87.10
	Precision	79.71	80.68	82.54	85.71
	Recall	82.21	84.71	84.65	88.55
XLM-R	F1	84.05	83.29	86.56	88.83
	Precision	82.68	80.55	84.82	87.35
	Recall	85.47	86.22	88.37	90.35
DistilBERT	F1	76.68	80.32	80.19	83.77
	Precision	76.50	80.81	79.91	83.81
	Recall	76.86	79.83	80.47	83.72

Table 3: Performance of mBERT, XML-R, DistilBERT and IgboBERT 2.0. We display the fine-tuned results of the models after 30 epoch at 1e-4 learning rate with varying dataset sizes.

4.4 Challenges of the Projection Method

We encountered some errors during the process of transferring the tags from English to Igbo sentences. This is as a result of the use of various orthographies during the translation of the parallel corpus. Some of these errors include:

- Missing entities and words in the Igbo translated text. For example, the sentence: "Chadwick Boseman who acted as T’Challa, king of Wakanda, who the movie ’Black Panther’ was made for, led other actors to Los Angeles where the ocassion was held." has this as the Igbo translation "duuru ndị ọzọ nọ n’ihe nkiri a gaa ebe emere mmeme a na los angeles" missing out the first part of the sentence ”Chadwick Boseman who acted as T’Challa, king of Wakanda, who the movie ’Black Panther’ was made for” in the translation.

- Variations in spelling of various Igbo words. For example, the word "Nigeria" has the following translations ("Nigeria","Naijiria", "Naija")},.
- Variations in translation of some words because of lack of standardized words for the Igbo words. The majority of these words are described in words in Igbo. For example, "8:30am" will be translated as "O jiri nkeji iri atọ wee gafe elekere asatọ nke ụtụtụ".

We addressed these challenges by adding multiple values of these word variations to one dictionary key in the mapping dictionary.

5 Conclusion and Future Work

This paper highlights the challenge of the lack of adequate datasets for building NLP models in general for low-resource languages with particular emphasis on the named entity recognition task for the Igbo language. A novel and efficient approach was proposed that takes advantage of parallel data in English - a well-resourced language with highly developed NER tools - and Igbo. This approach is based on the idea that the majority of name entities in the English text (mostly proper nouns) remain untranslated in the corresponding Igbo text. The concept of a mapping dictionary (see Figure 2) was introduced by creating a rule-based pipeline that enables the automatic transfer of the NER tags produced with the English NER tagger to the corresponding Igbo entities. With manual checks and corrections, this process not only speeds up the data creation effort but also produces the mapping dictionary which can serve as a key resource in other NLP tasks such as machine translation, and part-of-speech tagging. Our experiments show that model performance improves with increase in the data size when fine-tuned in the NER downstream task. Our IgboBERT 2.0 though was outperformed by the other models as shown in the tables but the performance is comparative when compared to the huge amount of data used in pre-training these SOTA models. We hope to improve our model further by creating further IgboNER datasets with the help of the NER mapping dictionary[11] created in this work.

Acknowledgements

We acknowledge the support of Tertiary Education Trust Fund (TETFund) in Nigeria, and would like to thank the anonymous reviewers for their useful comments and for reading the draft. We also acknowledge JW, Igbo-Radio and Kaoditaa who gave us permission to use data from their websites.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.298.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce

[11] https://github.com/Chiamakac/IgboNER-Models/tree/main/IgboNER$_2$.0

Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9: 1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

Maduabuchi Sennen Agbo. Orthography theories and the standard igbo orthography. 2013.

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. Building Named Entity Recognition Taggers via Parallel Corpora. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Alan Akbik and Roland Vollgraf. The projector: An interactive annotation projection visualization tool. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 43–48, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2008. URL https://aclanthology.org/D17-2008.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7415–7423, 2020.

Chiamaka Chukwuneke, Ignatius Ezeani, Paul Rayson, and Mahmoud El-Haj. IgboBERT models: Building and training transformer models for the Igbo language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5114–5122, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.547.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *ACL*, 2020a.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Maud Ehrmann, Marco Turchi, and Ralf Steinberger. Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 118–124, Hissar, Bulgaria, September 2011. Association for Computational Linguistics. URL `https://aclanthology.org/R11-1017`.

Jan Vium Enghoff, Søren Harrison, and Željko Agić. Low-resource named entity recognition via multi-source projection: Not quite there yet? In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 195–201, 2018.

Ignatius Ezeani, Paul Rayson, Ikechukwu E Onyenwe, Uchechukwu Chinedu, and Mark Hepple. Igbo-english machine translation: An evaluation benchmark. In *Eighth International Conference on Learning Representations*, 2020.

Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In *ACL*, 2020.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 6403–6416, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.478`.

Ralph Grishman and B Sundheim. Message understanding conference 6. In *Proceedings of the 16th conference on Computational linguistics*, volume 1, pp. 466, 1995.

Ruohao Guo and Dan Roth. Constrained labeled data generation for low-resource named entity recognition. In *FINDINGS*, 2021.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2580–2591, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.204. URL `https://www.aclweb.org/anthology/2020.emnlp-main.204`.

Michael A Hedderich, Lukas Lange, and Dietrich Klakow. Anea: distant supervision for low-resource named entity recognition. *arXiv preprint arXiv:2102.13129*, 2021.

William Hogan. An overview of distant supervision for relation extraction with a focus on denoising and pre-training methods. *ArXiv*, abs/2207.08286, 2022.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8066–8073, 2020.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 933–942, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.83. URL `https://aclanthology.org/2020.findings-emnlp.83`.

Bing Li, Yujie He, and Wenjin Xu. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*, 2021.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1054–1064, 2020.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10367–10378, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.810. URL `https://aclanthology.org/2021.emnlp-main.810`.

Mike D. Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Annual Meeting of the Association for Computational Linguistics*, 2009.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL `https://aclanthology.org/2020.findings-emnlp.195`.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.mrl-1.11`.

Louis Nnamdi Oraka. *The foundations of Igbo studies: A short history of the study of Igbo language and culture*. University Publishing Company, 1983.

Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1171–1182, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-1110`.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8118–8123, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.722. URL `https://aclanthology.org/2020.acl-main.722`.

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF, and Hugging Face. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. Denoising relation extraction from document-level distant supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3683–3688, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.300. URL `https://aclanthology.org/2020.emnlp-main.300`.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A Smith, and Jaime G Carbonell. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 369–379, 2018.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.